(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **08211895 A**

(43) Date of publication of application: **20 . 08 . 96**

(51) Int. Cl **G10L 3/02**
G10L 7/04
G10L 9/14
G10L 9/18

(21) Application number: **07295266**

(22) Date of filing: **14 . 11 . 95**

(30) Priority: **21 . 11 . 94 US 94 342494**

(71) Applicant: **ROCKWELL INTERNATL CORP**

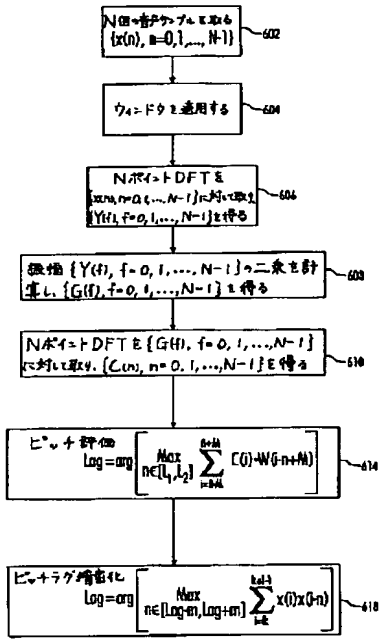
(72) Inventor: **SU HUAN-YU**
LI TOM HONG

(54) **SYSTEM AND METHOD TO EVALUATE PITCH LAG AND VOICE ENCODER**

(57) Abstract:

PROBLEM TO BE SOLVED: To provide a compact and accurate pitch evaluating system into which a multiple resolution analysis to encode a voice is incorporated.

SOLUTION: The pitch evaluating device and method evaluate pitch lag of a voice by using a multiple resolution system. This system contains a step of sampling a voice, a step of alternately applying discrete Fourier transform and a step of squaring a result. Next, DTF is performed to convert a voice sample into a separate area to squared amplitude. Next, an initial pitch lag is obtained with low resolution. After the pitch lag is evaluated with low resolution, algorithm made precise is on the basis of minimizing an anticipating error in a time area. Next, a pitch lag made preceise can be directly used in encoding a voice.



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-211895
(43)Date of publication of application : 20.08.1996

(51)Int.Cl. G10L 3/02
G10L 7/04
G10L 9/14
G10L 9/18

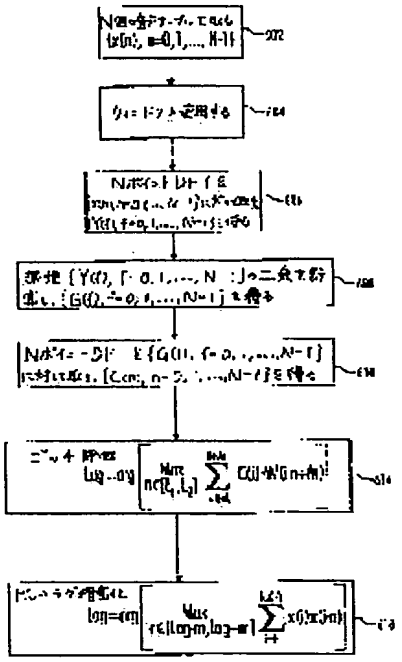
(21)Application number : 07-295266
(22)Date of filing : 14.11.1995

(71)Applicant : ROCKWELL INTERNATL CORP
(72)Inventor : SU HUAN-YU
LI TOM HONG

(30)Priority
Priority number : 94 342494 Priority date : 21.11.1994 Priority country : US

(54) SYSTEM AND METHOD TO EVALUATE PITCH LAG AND VOICE ENCODER

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a compact and accurate pitch evaluating system into which a multiple resolution analysis to encode a voice is incorporated.
SOLUTION: The pitch evaluating device and method evaluate pitch lag of a voice by using a multiple resolution system. This system contains a step of sampling a voice, a step of alternately applying discrete Fourier transform and a step of squaring a result. Next, DTF is performed to convert a voice sample into a separate area to squared amplitude. Next, an initial pitch lag is obtained with low resolution. After the pitch lag is evaluated with low resolution, algorithm made precise is on the basis of minimizing an anticipating error in a time area. Next, a pitch lag made preceise can be directly used in encoding a voice.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平8－211895

(43)公開日　平成８年(1996)８月20日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ１０Ｌ	3/02	Ｂ			
	7/04	Ｇ			
	9/14	Ｊ			
	9/18	Ｅ			

審査請求　未請求　請求項の数20　ＯＬ　（全 13 頁）

(21)出願番号　特願平7－295266

(22)出願日　平成７年(1995)11月14日

(31)優先権主張番号　０８／３４２４９４

(32)優先日　1994年11月21日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　590002448
ロックウェル・インターナショナル・コーポレイション
ROCKWELL INTERNATIONAL CORPORATION
アメリカ合衆国、90740－8250　カリフォルニア州、シール・ビーチ、シールビーチ・ブールバード、2201

(72)発明者　フアン－ユー・スー
アメリカ合衆国、92714　カリフォルニア州、アーバイン、ドュランゴ・アイル、398

(74)代理人　弁理士　深見　久郎　（外３名）

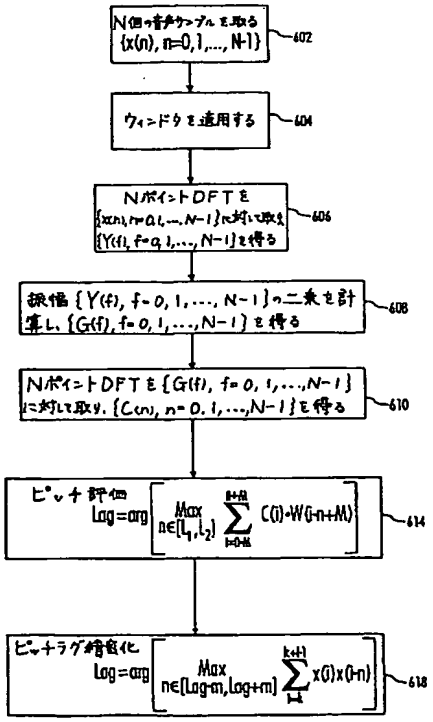
最終頁に続く

(54)【発明の名称】　ピッチラグを評価するためのシステムおよび方法、ならびに音声符号化装置および方法

(57)【要約】

【課題】　音声符号化のための多分解能分析を組入れる、簡潔かつ正確なピッチ評価システムを提供する。

【解決手段】　ピッチ評価装置および方法は多分解能方式を利用して音声のピッチラグを評価する。このシステムは音声をサンプルリングするステップと、交替に、離散フーリエ変換を適用するステップ、および結果を二乗するステップとを含む。次に、二乗された振幅に対してＤＦＴが音声サンプルを別の域へ変換するために行なわれる。初期のピッチラグは次に、低分解能で求められ得る。低分解能ピッチラグ評価を得た後で、精密化されたアルゴリズムが適用されて高分解能ピッチラグを得る。精密化されたアルゴリズムは時間域で予測誤差を最小にすることに基づく。精密化されたピッチラグを次に、音声符号化において直接用いることができる。



【特許請求の範囲】

【請求項１】　音声量子化および圧縮のためにピッチラグを評価するためのシステムであって、音声は複数の音声サンプルによって規定され、現在の音声サンプルの評価は過去のサンプルの線形結合に従う時間域で決定され、前記システムは、
音声サンプルに第１の離散フーリエ変換（ＤＦＴ）を適用するための手段を含み、前記第１のＤＦＴは関連した振幅を有し、さらに、
前記第１のＤＦＴの前記振幅を二乗するための手段と、
前記二乗された振幅に対して第２のＤＦＴを適用するための手段と、
時間域変換された音声サンプルに従って初期のピッチラグ値を決定するための手段と、
精密化されたピッチラグ値に従って前記音声サンプルを符号化するための手段とを含む、音声量子化および圧縮のためにピッチラグを評価するためのシステム。

【請求項２】　前記初期のピッチラグ値は関連した予測誤差を有し、前記システムは前記初期のピッチラグ値を精密化するための手段をさらに含み、前記関連した予測誤差は最小にされる、請求項１に記載のシステム。

【請求項３】　前記複数の音声サンプルを現在の符号化フレームへ分類するための手段と、
前記符号化フレームを複数のピッチサブフレームへ分割するための手段と、
前記ピッチサブフレームを複数の符号化サブフレームへ細分するための手段と、
前記現在の符号化フレームにおける各ピッチサブフレームの最後の符号化サブフレームに対して、それぞれラグ評価を表わす初期のピッチラグ評価lag_1およびlag_2を評価するための手段と、
先行の符号化フレームにおける第２のピッチサブフレームのピッチラグ評価lag_0を精密化するための手段と、
lag_1、lag_2、およびlag_0を線形的に補間し、前記符号化サブフレームのピッチラグ値を評価するための手段と、
各符号化サブフレームの補間されたピッチラグをさらに精密化するための手段とをさらに含む、請求項１に記載のシステム。

【請求項４】　少数のサンプルで概略的に表わすためにダウンサンプリング値へ前記音声サンプルをダウンサンプリングするための手段をさらに含む、請求項１に記載のシステム。

【請求項５】　前記初期のピッチラグ値は式（Lag_{scaled}＝音声サンプルの数／ダウンサンプリング値）によって補正される、請求項４に記載のシステム。

【請求項６】　前記初期のピッチラグ値を精密化するための前記手段は自己相関を含む、請求項１に記載のシステム。

【請求項７】　前記音声サンプルを受けるための音声入力手段と、
前記精密化されたピッチラグ値を処理し、符号化された音声として入力音声を再現するためのコンピュータと、
前記符号化された音声を出力するための音声出力手段とをさらに含む、請求項１に記載のシステム。

【請求項８】　入力音声を再現および符号化するための音声符号化装置であって、前記音声符号化装置は線形予測符号化（ＬＰＣ）パラメータと、音声を発生するために音声再現を誘起するように参照される複数のベクトルを表わす新規コードブックとを用いるようにされており、前記音声符号化装置は、
前記入力音声を受けるための音声入力手段と、
前記入力音声を処理するためのコンピュータとを含み、
前記コンピュータは、
前記入力音声内の現在の符号化フレームを切出すための手段と、
前記符号化フレームを複数のピッチサブフレームへ分割するための手段と、
Ｎ個の音声サンプルを有するピッチ分析ウインドウを規定するための手段とを含み、前記ピッチ分析ウインドウは前記ピッチサブフレームに対して延び、前記コンピュータは、
各ピッチサブフレームに対して初期のピッチラグ値を評価するための手段と、
各ピッチサブフレームを複数の符号化サブフレームへ分割するための手段とを含み、各ピッチサブフレームに対する前記初期のピッチラグ評価は、前記現在の符号化フレームにおける各ピッチサブフレームの最後の符号化サブフレームに対するラグ評価を表わし、前記コンピュータは、
前記評価されたピッチラグ値を前記ピッチサブフレームの間で線形的に補間し、各符号化サブフレームに対してピッチラグ評価を決定するための手段と、
各符号化サブフレームの前記線形的に補間されたラグ値を精密化するための手段とを含み、前記装置はさらに、
前記精密化されたピッチラグ値に従って再現された音声を出力するための音声出力手段を含む、入力音声を再現および符号化するための音声符号化装置。

【請求項９】　前記コンピュータは、
少数のサンプルで表わすためにダウンサンプリング値Ｘへ前記Ｎ個の音声サンプルをダウンサンプリングするための手段と、
補正されたラグ値Lag_{scaled}＝Ｎ／Ｘであるように前記ピッチラグ値を補正するための手段とをさらに含む、請求項８に記載の装置。

【請求項１０】　サンプリング速度Ｒで前記入力音声をサンプリングするサンプリング手段をさらに含み、前記Ｎ個の音声サンプルは式Ｎ＝Ｒ＊Ｘに従って決定される、請求項８に記載の装置。

【請求項１１】　Ｘ＝２５ｍｓ、Ｒ＝８０００Ｈｚ、およびＮ＝３２０サンプルである、請求項１０に記載の装置。

【請求項１２】　各符号化フレームはおおよそ４０ｍｓの長さを有する、請求項８に記載の装置。

【請求項１３】　音声量子化および圧縮のためにピッチラグを評価するための方法であって、前記音声は複数の音声サンプルによって規定され、現在の音声サンプルの評価は過去のサンプルの線形結合に従う時間域で決定され、前記方法は、
音声サンプルに第１の離散フーリエ変換（ＤＦＴ）を適用するステップを含み、前記第１のＤＦＴは関連した振幅を有し、さらに、
前記第１のＤＦＴの振幅を二乗するステップと、
前記第１のＤＦＴの前記二乗された振幅に対して第２のＤＦＴを適用するステップと、
時間域変換された音声サンプルに従って初期のピッチラグ値を決定するステップとを含み、前記初期のピッチラグ値は関連した予測誤差を有し、前記方法はさらに、
自己相関を用いて前記初期のピッチラグ値を精密化するステップを含み、前記関連した予測誤差は最小にされ、さらに、
前記精密化されたピッチラグ値に従って前記音声サンプルを符号化するステップを含む、音声量子化および圧縮のためにピッチラグを評価するための方法。

【請求項１４】　前記複数の音声サンプルを現在の符号化フレームへ分類するステップと、
前記符号化フレームを複数のピッチサブフレームへ分割するステップと、
前記ピッチサブフレームを複数の符号化サブフレームへ細分するステップと、
前記現在の符号化フレームにおける各ピッチサブフレームの最後の符号化サブフレームに対して、それぞれラグ評価を表わす初期のピッチラグ評価 lag_1 および lag_2 をそれぞれ評価するステップと、
先行の符号化フレームにおける第２のピッチサブフレームのピッチラグ評価 lag_0 を精密化するステップと、
lag_1、lag_2、および lag_0 を線形的に補間し、前記符号化サブフレームのピッチラグ値を評価するステップと、
各符号化サブフレームの補間されたピッチラグをさらに精密化するステップとをさらに含む、請求項１３に記載の方法。

【請求項１５】　少数のサンプルで概略的に表わすためにダウンサンプリング値へ前記音声サンプルをダウンサンプリングするステップをさらに含む、請求項１３に記載の方法。

【請求項１６】　式（Lag_{scaled}＝音声サンプルの数／ダウンサンプリング値）に従って前記初期のピッチラグ値を補正するステップをさらに含む、請求項１５に記載の方法。

【請求項１７】　前記音声サンプルを受けるステップと、
前記精密化されたピッチラグ値を処理し、符号化された音声として前記入力音声を再現するステップと、
前記符号化された音声を出力するステップとをさらに含む、請求項１３に記載のシステム。

【請求項１８】　入力音声を再現および符合化するための音声符号化方法であって、音声符号化装置は線形予測符号化（ＬＰＣ）パラメータと、音声を発生するために音声再現を誘起するように参照される複数のベクトルを形成する擬似ランダム信号を表わす新規コードブックとを用いるようにされており、前記音声符号化方法は、
前記入力音声を受取り、処理するステップと、
前記入力音声を処理するステップとを含み、前記処理するステップは、
前記入力音声内で音声符号化フレームを決定するステップと、
前記符号化フレームを複数のピッチサブフレームへ細分するステップと、
Ｎ個の音声サンプルを有するピッチ分析ウインドウを規定するステップとを含み、前記ピッチ分析ウインドウは前記ピッチサブフレームにわたって延び、前記処理するステップは、
各ピッチサブフレームに対して初期のピッチラグ価を概略的に評価するステップと、
各ピッチサブフレームに対する初期のピッチラグ評価が各ピッチサブフレームの最後の符号化サブフレームに対するラグ評価を表わすように、各ピッチサブフレームを複数の符号化サブフレームへ分割するステップと、
前記評価されたピッチラグ値を前記ピッチサブフレームの間で補間し、各符号化サブフレームに対してピッチラグ評価を決定するステップと、
線形的に補間されたラグ値を精密化するステップとを含み、前記方法はさらに、
精密化されたピッチラグ値に従って再現された音声を出力するステップを含む、入力音声を再現および符号化するための音声符号化方法。

【請求項１９】　前記処理するステップは、
少数のサンプルで表わすためにダウンサンプリング値Ｘへ前記Ｎ個の音声サンプルをダウンサンプルするステップと、
補正されたラグ値 Lag_{scaled}＝Ｎ／Ｘであるように前記ピッチラグ値を補正するステップとをさらに含む、請求項１８に記載の装置。

【請求項２０】　前記Ｎ個の音声サンプルが式Ｎ＝Ｒ＊Ｘに従って決定されるように、サンプリング速度Ｒで前記入力音声をサンプリングするステップをさらに含む、請求項１８に記載の方法。

【発明の詳細な説明】

【０００１】
【発明の背景】信号のモデル化およびパラメータ評価はデータ圧縮、復元、および符号化においてますます重要な役割を果たす。基本的な話声音をモデル化するために、音声信号は離散波形としてサンプリングされ、デジタル的に処理されなければならない。線形予測符号化（ＬＰＣ）と称されるあるタイプの信号符号化技術において、何らかの特定の時間指標での信号値は前の値の線形関数としてモデル化される。こうして、後の信号はこれまでの値に従って線形的に予測される。結果として、効果的な信号表現は信号を表わすために、ある予測パラメータを評価し、かつ適用することによって決定できる。現在、符号励起線形予測（ＣＥＬＰ）を含む音声符号化のためにＬＰＣ技術が用いられている。
【０００２】ピッチ情報は符号化の目的に対して、確かな音の指標および表示であると認められている。ピッチは話者の音声の基本的な特徴またはパラメータを記述する。人間の音声は一般に容易には数学的に定量化できないので、音声のピッチデータを効果的に評価できる音声評価モデルが、より的確かつ正確に符号化され、かつ復号された音声を提供する。しかしながら、あるＣＥＬＰ（たとえば、ベクトル和励起線形予測（ＶＳＥＬＰ）、マルチパルス、正規パルス、代数的ＣＥＬＰなど）およびＭＢＥコーダ／デコーダ（「コーデック」）のような現在の音声符号化モデルにおいて、ピッチ評価アルゴリズムは正確さが高く、かつ複雑さが低いことを必要とするために、ピッチ評価がしばしば困難である。
【０００３】いくつかのピッチラグ評価方式は上述されたコーデック（時間域方式、周波数域方式、およびケプストラム域方式）と関連して用いられる。ピッチラグおよび音声再現の間に密接な関係があるために、ピッチ評価の正確さは通話品質に直接的な影響を及ぼす。ＣＥＬＰコーダでは、音声発生は予測（長期ピッチ予測および短期線形予測）に基づく。図１は典型的なＣＥＬＰコーダによる音声再生のブロック図を示す。
【０００４】音声データを圧縮するためには、不可欠な情報のみを抽出して冗長の伝送を回避することが望ましい。音声は短いブロックに分類でき、ここで代表的なパラメータはあらゆるブロックにおいて識別できる。図１に示されるように、ＣＥＬＰ音声コーダは良質な音声を発生するために、ＬＰＣパラメータ１１０と、（ラグおよびその係数を含む）ピッチラグパラメータ１１２と、その利得パラメータ１１６を備える最適な新規コードベクトル１１４とを符号化されるべき入力音声から抽出しなければならない。コーダは適当な符号化方式を実現することによってＬＰＣパラメータを量子化する。各パラメータの量子化の指標は、音声デコーダに記憶または伝送されるべき情報を含む。ＣＥＬＰコーデックでは、ピッチ予測パラメータ（ピッチラグおよびピッチ係数）の決定は時間域で行なわれるが、ＭＢＥコーデックでは、ピッチパラメータは周波数域で評価される。
【０００５】ＬＰＣ分析の後でＣＥＬＰエンコーダは、（通例約１０－４０ｍｓで取られる）現在の音声符号化フレームのために適当なＬＰＣフィルタ１１０を決定する。ＬＰＣフィルタは次式によって表わされる。
【０００６】
【数１】

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_{np} z^{-np}$$

またはｎ番目のサンプルが次式によって予測できる

$$y(n) = \sum_{k=1}^{np} a_k * y(n-k)$$

【０００７】この式において、ｎｐはＬＰＣ予測次数（通例、約１０）であり、ｙ（ｎ）はサンプリングされた音声データであり、ｎは時間指標を表わす。上記のＬＰＣの式は、過去のサンプルの線形結合に従って現在のサンプルの評価を記述する。人間の耳の感度をモデルとするＬＰＣフィルタに基づく聴感補正フィルタはここで次式によって規定される。
【０００８】
【数２】

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_3)} \qquad \text{但し} \quad 0 < \gamma_2 < \gamma_1 \leq 1$$

【０００９】所望のピッチパラメータを抽出するために、次の重み付き符号化誤差エネルギを最小にするピッチパラメータは各符号化サブフレームについて計算されなければならず、ここで１つの符号化フレームは、分析および符号化のためにいくつかの符号化サブフレームへ分割できる。
【００１０】
【数３】

$$d = \|T - \beta P_{Lag} H - \alpha C_i H\|^2$$

【００１１】この式において、Ｔは知覚的にフィルタされた入力信号を表わす目標信号であり、ＨはフィルタＷ（ｚ）／Ａ（ｚ）のインパルス応答行列である。P_{Lag}はピッチラグ「Ｌａｇ」と、所定のラグについて独自に規定された予測係数βとを有するピッチ予測寄与であり、C_iはコードブックにおける指標ｉおよびその対応する利得αに関連したコードブック寄与である。典型的には、人間の音声のピッチは２ｍｓから２０ｍｓの間で異なる。したがって、音声が８ＫＨｚのサンプリング速

度でサンプリングされると、ピッチラグは概算で２０サンプルから１４７サンプルに対応する。さらに、ｉは０およびＮｃ－１の間の値を取り、ここでＮｃは新規コードブックのサイズである。

【００１２】１タップピッチ予測子および１つの新規コードブックを想定する。しかしながら、典型的にピッチ予測子の一般的な形状は多タップ方式であり、新規コードブックの一般的な形状は多レベルベクトル量子化であるか、または、複数の新規コードブックを用いる。より詳細には、音声の符号化において、１タップピッチ予測子は現在の音声サンプルが１つの過去の音声サンプルによって予測できることを示し、一方多タップ予測子は現

$$\underset{Lag\in[L_1,L_2]}{Max}\ \frac{\left(TH^T P_{Lag}^T\right)^2}{\left\|P_{Lag}H\right\|^2}$$

式(１)

【００１５】この時間域方式は真のピッチラグを決定できるが、高いピッチ周波数を有する女性の音声には、式（１）によって求められるピッチラグは真のラグではなく、真のラグの倍数となり得る。この評価誤差を回避するために、評価誤差を訂正（たとえば、ラグの平滑化）する付加的なプロセスが必要であり、これはそれと引換えに不所望な複雑さを引起こす。

【００１６】しかしながら、このように過度に複雑であることは、時間域方式を用いる際の著しい欠点である。たとえば、整数のラグのみを用いてラグを決定するために、時間域方式は１秒当り３００万回の動作（３ＭＯＰ）を少なくとも必要とする。さらに、ピッチラグの平滑化および分数のピッチラグが用いられるならば、複雑さはほぼ４ＭＯＰであろう。実際には、容認可能な正確さでフルレンジのピッチラグ評価を実行するために、概算で１秒当り６００万回のデジタル信号処理機械命令（６ＤＳＰ　ＭＩＰ）が必要とされる。したがって、ピッチ評価は４から６のＤＳＰ　ＭＩＰを必要とすると一般に認められている。ピッチ評価の複雑さを減少できる方式は他にもあるが、そのような方式はしばしば質を犠牲にする。

【００１７】正弦コーダの類で重要な要素であるＭＢＥコーダでは、符号化パラメータは周波数域において抽出され、かつ量子化される。ＭＢＥ音声モデルは図２から図４に示される。図２および図３に説明されるＭＢＥ音声エンコーダ／デコーダ（「ボゴーダ」）では、基本周波数（またはピッチラグ）２１０、有声／無声決定２１２、およびスペクトルエンベロープ２１４は周波数域において入力音声から抽出される。パラメータは次に、記憶または伝送できるビットストリームへ量子化され、かつ符号化される。

【００１８】ＭＢＥボコーダでは、良質な音声を達成するために基本周波数が高い正確さで評価されなければならない。基本周波数の評価は２段階で行なわれる。第１在の音声サンプルが複数の過去の音声サンプルによって予測できることを意味する。

【００１３】複雑さについて懸念があるために、最適な方式に準ずる方式が音声符号化方式において用いられてきた。たとえば、ピッチラグ評価は２．５ｍｓから１８．５ｍｓをカバーするために、L_1 およびL_2 サンプルの間の範囲で起こり得るラグ値を単に評価することによってなされてもよい。したがって、評価されたピッチラグ値は次式を最大にすることによって決定される。

【００１４】

【数４】

に、初期のピッチラグが２１サンプルから１１４サンプルの範囲内で探索され、周波数域において入力音声２１６および合成された音声２１８の間で重み付き平均二乗誤差方程式３１０（図３）を最小にすることによって８０００Ｈｚのサンプリング速度で２．６ｍｓから１４．２５ｍｓをカバーする。元の音声および合成された音声の間の平均二乗誤差は次式によって与えられる。

【００１９】

【数５】

$$E=\frac{1}{2\pi}\int_{-\pi}^{\pi}G(\omega)\left|s(\omega)-\hat{S}(\omega)\right|d\omega$$

【００２０】この式において、Ｓ（ω）は元の音声スペクトルであり、Ｓ^（ω）（^は大文字Ｓの上にあると見なされる）は合成された音声スペクトルであり、Ｇ（ω）は周波数依存重み付き関数である。図４に示されるように、ピッチ追跡アルゴリズム４１０は、隣接するフレームのピッチ情報を用いることによって、初期のピッチラグ評価４１２を更新するのに用いられる。

【００２１】この方式を用いるのは、基本周波数が隣接するフレームの間で不意には変化するはずはないという仮定のためである。２つの過去の隣接するフレームおよび２つの未来の隣接するフレームのピッチ評価はピッチ追跡のために使用される。次に、（２つの過去のフレームおよび２つの未来のフレームを含む）平均二乗誤差は最小にされて現在のフレームの新しいピッチラグ値を求める。初期のピッチラグを追跡した後で、ピッチラグ多重検査方式４１４が多重ピッチラグを除去するために適用され、ピッチラグを平滑化する。

【００２２】図４を参照すると、基本周波数評価の第２段階でピッチラグ精密化４１６が用いられてピッチ評価の正確さを高める。ピッチラグ候補値は初期のピッチラグ評価に基づいて形成される（すなわち、新しいピッチラグ候補値は、初期のピッチラグ評価からある分数を加

算し、または減算することによって形成される）。したがって、精密化されたピッチラグ評価４１８は、平均二乗誤差関数を最小にすることによってピッチラグ候補の中で決定できる。

【００２３】しかしながら、周波数域ピッチ評価はある欠点を有する。第１に、複雑さが非常に高い。第２に、ピッチラグは２．５ｍｓから１４．２５ｍｓしかカバーしない２０および１１４サンプルの範囲内で探索されて、２５６ポイントＦＦＴに対処するように２５６サンプルにウインドウサイズを制限しなければならない。しかしながら、非常に低いピッチ周波数の話者には、または１４．２５ｍｓを超えるピッチラグを有する音声には、２５６サンプルウインドウ内で十分な数のサンプルを集めるのが不可能である。さらに、音声フレームにわたって評価されるのは、平均されたピッチラグだけである。

【００２４】１９６７年にエイ．エム．ノル（A.M.Noll）によって提案されたケプストラム域ピッチラグ評価（図５）を用いて、変形された方法が他に提案された。ケプストラム域ピッチラグ評価では、５１０でおおよそ３７ｍｓの音声がサンプリングされるので、可能な最大のピッチラグ（たとえば、１８．５ｍｓ）の少なくとも２周期がカバーされる。次に、５１２ポイントＦＦＴは音声フレームウインドウに（ブロック５１２で）適用され、周波数スペクトルを獲得する。周波数スペクトルの対数５１４の振幅を取って、別の５１２ポイント逆ＦＦＴ５１６がケプストラムを得るために適用される。重み付け関数５１８はケプストラムに適用され、ケプストラムのピークはピッチラグを決定するために５２０で検出される。次に、追跡アルゴリズム５２２が実行されて、いかなるピッチ倍数をも除去する。

【００２５】しかしながら、ケプストラムピッチ検出方法にはいくつかの欠点が見受けられる。たとえば、計算上の要求が高い。８ＫＨｚのサンプリング速度において２０サンプルおよび１４７サンプルの間でピッチの範囲をカバーするために、５１２ポイントＦＦＴは二度行なわれなければならない。ケプストラムピッチ評価が分析フレームにわたる平均されたピッチラグの評価のみを提供するので、評価の正確さが不十分である。しかしながら、低ビット転送速度音声符号化については、ピッチラグ値が短い時間期間にわたって評価されることは重要である。結果として、ケプストラムピッチ評価は今日、高質な低ビット転送速度音声符号化についてはほとんど用いられない。したがって、上述された方式の各々に制限があるために、効果的なピッチラグ評価のための手段には、高質な低ビット転送速度音声符号化の必要を満たすことが所望される。

【００２６】

【発明の概要】したがって、この発明の目的は、複雑さが最小であって正確さが高いことを必要とする、音声符号化のための多分解能分析を組入れるピッチ評価システムを提供することである。特定の実施例では、この発明はＣＥＬＰ技術ならびにさまざまな他の音声符号化および認識システムを用いる音声符号化の装置および方法を対象とする。したがって、必要な高い正確さを維持しながら、よりよい結果がより少ない計算手段でもたらされる。

【００２７】これらの目的および他の目的は、この発明の実施例に従って、音声の的確な再現および再生を速くかつ効果的に可能にするピッチラグ評価方式によって達成される。ピッチラグは所定の音声フレームについて抽出され、次に、各サブフレームについて精密化される。最小の数の音声サンプルが音声を直接サンプリングすることによって獲得された後で、離散フーリエ変換（ＤＦＴ）が適用され、結果として生じる振幅が二乗される。第２のＤＦＴが次に行なわれる。したがって、フレーム内の音声サンプルに対する的確な初期のピッチラグは、８ＫＨｚのサンプリング速度で２０サンプルの可能な最小値と１４７サンプルの最大ラグ値との間で決定できる。初期のピッチラグ評価を獲得した後で、時間域精密化がさらに評価の正確さを向上するために各サブフレームについて行なわれなければならない。

【００２８】

【好ましい実施例の詳細な説明】この発明の好ましい実施例に従ったピッチラグ評価方式は一般に図６、７、８および９において示される。まず、Ｎ個の音声サンプル｛ｘ（ｎ），ｎ＝０，１，…，Ｎ－１｝が集められる。（図６のステップ６０２）Ｎはたとえば、８０００Ｈｚのサンプリング速度で典型的な４０ｍｓの音声ウインドウに対処するために３２０個の音声サンプルに等しくてもよい。Ｎの値はおおまかに評価された音声周期によって決定され、ここで少なくとも２周期が音声スペクトルを発生するために一般に必要とされる。このように、Ｎが可能な最大のピッチラグの２倍よりも大きくなくてはならず、ここでは｛ｘ（ｎ），ｎ＝０，１，…，Ｎ－１｝である。さらに、少なくとも２ピッチ周期をカバーするハミングウインドウ６０４または他のウインドウが好ましくは実現される。

【００２９】

【数６】

NポイントDFTがステップ606で$\{x(n), n=0,1,\ldots,N-1\}$にわたって適用され、振幅$\{Y(f), f=0,1,\ldots,N-1\}$を得る。ここで

$$Y(f)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nf/N} \quad \text{但し} \quad f=0,1,\ldots,N-1 \qquad \text{式(2)}$$

次に、Y（f）はステップ608で次式に従って二乗される。

$$G(f)=|Y(f)|^2 \quad \text{但し} \quad f=0,1,\ldots,N-1 \qquad \text{式(3)}$$

第2のNポイントDFTがステップ610でG（f）に適用されて以下を獲得する。

$$C(n)=\sum_{f=0}^{N-1}G(f)e^{-j2\pi nf/N} \quad \text{但し} \quad n=0,1,\ldots,N-1 \qquad \text{式(4)}$$

【0030】この発明の実施例に従って、関数G（f）ではなくG（f）の対数が式（4）において用いられる従来のケプストラム変換とC（n）とが異なることが認識される。この違いの原因は一般的には複雑さである。除去されなければ実質的により多くの計算資源を必要とする対数関数を除去することによって複雑さを減少することが望ましい。さらに、ケプストラムまたはC（n）関数を用いるピッチラグ評価方式を比較すると、音声の無声または遷移区間のみに対して異なった結果が獲得されていた。たとえば、無声または遷移音声に対してピッチの定義が不明確である。遷移音声にはピッチがないといわれてきたが、誤差を最小にするために何らかの予測を常に示すことができるともいわれる。

【0031】したがって、一旦C（n）が決定されると（ステップ610）、所定の音声フレームに対するピッチラグは次式を解くことによってステップ614で求められ得る。

【0032】

【数7】

$$Lag=arg\left[\underset{n\in[L_1,L_2]}{Max}\sum_{i=n-M}^{n+M}C(i)\bullet W(i-n+M)\right] \qquad \text{式(5)}$$

【0033】この式において、arg［・］は内部最適化関数を満たす変数nを決定し、L_1およびL_2はそれぞれ可能な最小のピッチラグおよび可能な最大のピッチラグとして規定される。音声符号化の便宜上、L_2およびL_1の間の差は2進数表現のために2の累乗であることが望ましい。好ましい実施例では、L_1およびL_2はそれぞれ20および147の値を取って典型的な人間の音声のピッチラグ範囲の2．5msから18．375msをカバーし、ここでL_1およびL_2の間の間隔は2の累乗である。W（i）は重み付き関数であり、2M＋1はウインドウサイズを表わす。好ましくは、｛W（i）＝1，i＝0，1，…，2M｝であり、M＝1である。

【0034】結果として生じるピッチラグは平均された値であるが、それは信頼できて的確であるということがわかった。平均化から生じる効果は絶対的に大きな分析ウインドウサイズによるものであり、147サンプルのラグに対して、ウインドウサイズはラグ値の少なくとも2倍であるべきである。しかしながら、不所望なことに、典型的に小さいピッチラグを示す女性の話者のようなある話者からの信号は、このような大きなウインドウでは4から10ピッチ周期を含み得る。ピッチラグに変化があれば、提案されたピッチラグ評価は平均されたピッチラグしか生成しない。結果として、音声符号化においてこのような平均されたピッチラグを用いることで音声評価および再生に大きな劣化が生じ得る。

【0035】音声におけるピッチ情報の相対的に速い変化のために、CELPモデルに基づくほとんどの音声符号化システムはサブフレームごとに一度ピッチラグを評価し、かつ伝送する。こうして、典型的には2msから10msの長さ（16から80サンプル）であるいくつかの音声サブフレームへ1つの音声フレームが分割されるCELP型音声符号化において、ピッチ情報は各サブフレームで更新される。したがって、正確なピッチラグ値はサブフレームのためにのみ必要とされる。しかしながら、上記の方式に従って評価されたピッチラグは、平均化から生じる影響のために正確な音声符号化には十分な正確さを有さない。

【0036】こうして、この発明の特定の実施例において、評価の正確さを向上させるために、初期のピッチラグ評価に基づいた精密化された探索が時間域において行

なわれる（ステップ６１８）。簡単な自己相関方法がほぼ平均されたＬａｇ値で特定の符号化周期またはサブフレームに対して行なわれる。

【００３７】
【数８】

$$Lag = arg\left[\underset{n\in[Lag-m,Lag+m]}{Max}\sum_{i=k}^{k+l-1} x(i)x(i-n)\right] \qquad 式(6)$$

【００３８】この式において、ａｒｇ［・］は内部最適化関数を満たす変数ｎを決定し、ｋはサブフレームの第１のサンプルを示し、ｌは精密化ウインドウサイズを表わし、ｍは探索範囲である。的確なピッチラグ値を決定するために、精密化ウインドウサイズは少なくとも１ピッチ周期であるべきである。しかしながら、ウインドウは平均化の影響を避けるためにあまりに大きすぎてはならない。たとえば、好ましくはｌ＝Ｌａｇ＋１０、およびｍ＝５である。こうして、式（６）の時間域精密化に従って、より正確なピッチラグが評価されてサブフレームの符号化に適用できる。

【００３９】動作時において、高速フーリエ変換（ＦＦＴ）が一般的なＤＦＴよりも計算上効果的である場合もあるが、ＦＦＴを用いる際の欠点はウインドウサイズが２の累乗でなければならないことである。たとえば、１４７サンプルの最大のピッチラグは２の累乗ではないことが示されてきた。最大のピッチラグを含むためには、５１２サンプルのウインドウサイズが必要である。しかしながら、このことで、上述された平均化から生じる影響のために女性の音声に対するピッチラグ評価の質が悪くなり、多量の計算が必要となる。２５６サンプルのウインドウサイズが用いられるならば、平均化から生じる影響は減少され、複雑さが一層少なくなる。しかしながら、このようなウインドウを用いると音声中の１２８サンプルよりも大きなピッチラグには対処できない。

【００４０】これらの問題のいくつかを克服するために、この発明の代替の好ましい実施例は２５６ポイントＦＦＴを利用して複雑さを減少し、変更された信号を用いてピッチラグを評価する。信号を変更するのはダウンサンプリングプロセスである。図７および図８を参照すると、Ｎ個の音声サンプル｛ｘ（ｎ），ｎ＝０，１，…，Ｎ－１｝が集められ（ステップ７０２）、Ｎは最大のピッチラグの２倍よりも大きい。次に、Ｎ個の音声サンプルが次式に従って、線形補間を用いて２５６個の新しい分析サンプルへダウンサンプリングされる（ステップ７０４）。

【００４１】
【数９】

$$y(i) = x([i\bullet\lambda]) + \{x([i\bullet\lambda]+1) - x([i\bullet\lambda])\}(i\bullet\lambda - [i\bullet\lambda])$$
$$\text{但し}\quad i = 0, 1, \ldots, 255$$

【００４２】この式において、λ＝Ｎ／２５６であり、角括弧内の値すなわち［ｉ・λ］はｉ・λ以下の最大の整数値を示す。次に、ステップ７０５でハミングウインドウまたは他のウインドウが補間されたデータに適用される。

【００４３】ステップ７０６では、ピッチラグ評価は２５６ポイントＦＦＴを用いてｙ（ｉ）にわたって行なわれ、振幅Ｙ（ｆ）を発生する。次に、ステップ７０８からステップ７１０は図６に関して説明されたのと同様に実行される。しかしながら、Ｇ（ｆ）はさらにフィルタされ（ステップ７０９）、ピッチ検出のためには有用ではない、Ｇ（ｆ）の高周波数成分を減少する。一旦ｙ（ｉ）のラグすなわちLag_yが式（５）に従って求められれば（ステップ７１４）、これはステップ７１６で再スケールされてピッチラグ評価を決定する。

【００４４】
【数１０】

$$Lag = Lag_y \bullet \lambda$$

【００４５】要約すると、符号化フレームのための初期のピッチ評価を求める上記の手順は以下のとおりである。

【００４６】（１）　標準４０ｍｓの符号化フレームをピッチサブフレーム８０２および８０４へ細分する。各ピッチサブフレームはおおよそ２０ｍｓの長さである。

【００４７】（２）　ピッチ分析ウインドウ８０６が最後のサブフレームの中心に位置決めされるようにＮ＝３２０個の音声サンプルを取り、提案されたアルゴリズムを用いてそのサブフレームに対するラグを求める。

【００４８】（３）　ピッチサブフレームに対する初期のピッチラグ値を決定する。

次に、時間域精密化が元の音声サンプルｘ（ｎ）にわたってステップ７１８で行なわれる。こうして、この発明の実施例において、複雑さを減少してなお、高い正確さを維持しながらピッチラグ値が的確に評価できる。この発明のＦＦＴ実施例を用いると、１２０よりも大きいピッチラグ値に対処するのは困難ではない。

【００４９】より詳細には、時間域精密化は元の音声サンプルにわたって行なわれる。たとえば、４０ｍｓの符号化フレームは図９に示されるようにまず、８個の５ｍｓのサブフレーム８０８へ分割される。初期のピッチラグ評価lag_1およびlag_2は、現在の符号化フレームにおける各ピッチサブフレームの最後の符号化サブフレームに対するラグ評価である。lag_0は先行の符号化フレームにおける第２のピッチサブフレームの精密化

されたラグ評価である。lag_1、lag_2、およびlag_0の間の関係は図９に示される。

【００５０】初期のピッチラグlag_1およびlag_2は次式に従って最初に精密化されて、その正確さを向上させる（図８のステップ７１８）。

【００５１】

【数１１】

$$lag_i = arg\left[\max_{n\in[lag_i-M,lag_i+M]}\sum_{k=N_i}^{N_i+L-1} x(k)\bullet x(k-n)\right]$$

但し　$i=1,2$

【００５２】ここでN_iは、ピッチlag_iに対するピッチサブフレームにおける開始サンプルの指標である。好ましくは、Mは１０と選択され、Lはlag_i＋１０であり、iはピッチサブフレームの指標を示す。

【００５３】一旦初期のピッチラグの精密化が完了すると、符号化サブフレームのピッチラグが決定できる。符号化サブフレームのピッチラグはlag_1、lag_2、およびlag_0を線形的に補間することによって評価される。符号化サブフレームのピッチラグ評価の正確さは、次の手順に従って各符号化サブフレームの補間されたピッチラグを精密化することによって向上する。精密化された初期のピッチ評価lag_1、lag_2、およびlag_0に基づく符号化サブフレームの補間されたピッチラグを｛lag_I（ｉ），ｉ＝０，１，…，７｝が表わす場合、lag_I（ｉ）は次式によって決定される。

【００５４】

【数１２】

$$lag_I(i)=\begin{cases} lag_0+(lag_1-lag_0)*\dfrac{i+1}{4} & \text{但し}\quad i=0,1,2,3 \\ lag_1+(lag_2-lag_1)*\dfrac{i-3}{4} & \text{但し}\quad i=4,5,6,7 \end{cases}$$

【００５５】線形補間によって与えられるピッチラグ評価の正確さが十分ではないので、さらなる改良が必要となるであろう。与えられたピッチラグ評価｛lag_I（ｉ），ｉ＝０，１，…，７｝に対して、各lag_I（ｉ）は次式によってさらに精密化される（ステップ７２２）。

【００５６】

【数１３】

$$lag(i) = arg\left[\max_{n\in[lag_I(i)-M,lag_I(i)+M]}\sum_{k=N_i}^{N_i+L-1} x(k)\bullet x(k-n)\right]$$

但し　$i=0,1,\ldots,7$

【００５７】ここでＮｉはピッチｌａｇ（ｉ）に対する符号化サブフレームにおける開始サンプルの指標である。例では、Mは３と選択され、Ｌは４０に等しい。

【００５８】さらに、ピッチラグの線形補間は音声の無声区間において重要である。何らかの分析方法によって求められたピッチラグは無声音声に任意に配分される傾向を有する。しかしながら、相対的に大きいピッチサブフレームサイズのために、各サブフレームに対するラグが（上の手順（２）で求められる）始めに決定されたサブフレームラグにあまりにも近い場合、元々は音声にはなかった不所望な人工の周期性が加えられる。さらに線形補間は、質の悪い無声音声に関連した問題を簡単に解決する。さらに、サブフレームのラグは任意である傾向を有するので、各サブフレームに対するラグは一旦補間されると、これも非常に任意に配分され、このことが音声の質を保証する。

【図面の簡単な説明】

【図１】ＣＥＬＰ音声モデルのブロック図である。

【図２】ＭＢＥ音声モデルのブロック図である。

【図３】ＭＢＥエンコーダのブロック図である。

【図４】ＭＢＥボコーダにおけるピッチラグ評価のブロック図である。

【図５】ケプストラムに基づくピッチラグ検出方式のブロック図である。

【図６】この発明の実施例に従うピッチラグ評価の動作上のフロー図である。

【図７】この発明の別の実施例に従うピッチラグ評価のフロー図である。

【図８】この発明の別の実施例に従うピッチラグ評価のフロー図である。

【図９】図６の実施例に従う音声符号化の図である。

【符号の説明】

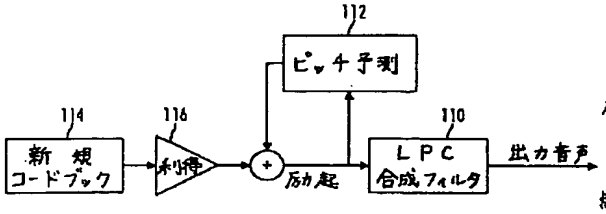
８０２　ピッチサブフレーム

８０４　ピッチサブフレーム

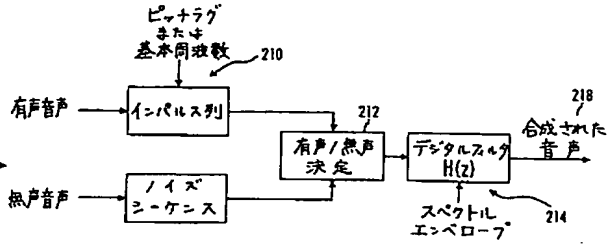
８０６　ピッチ分析ウインドウ

８０８　サブフレーム

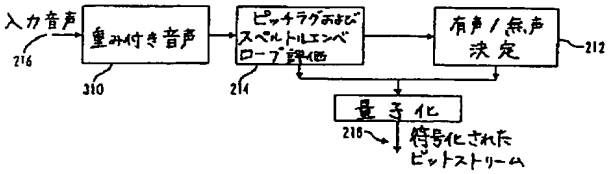
【図１】



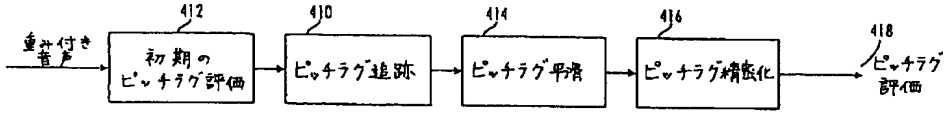
【図２】



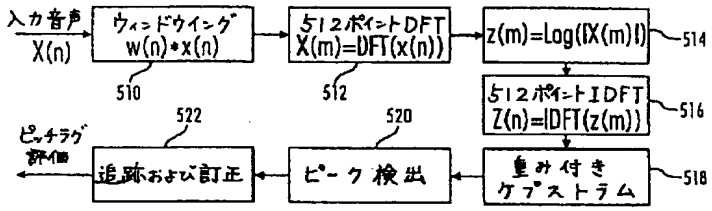
【図３】



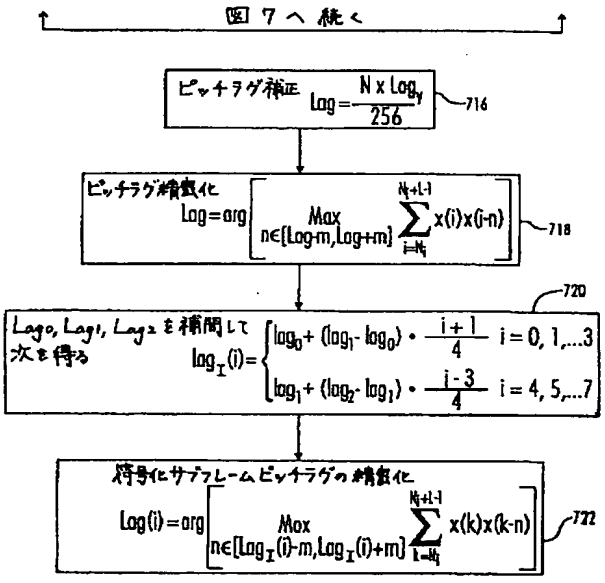
【図４】



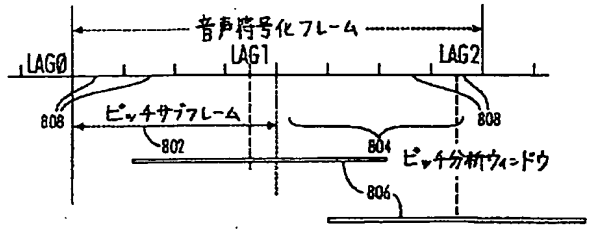
【図５】



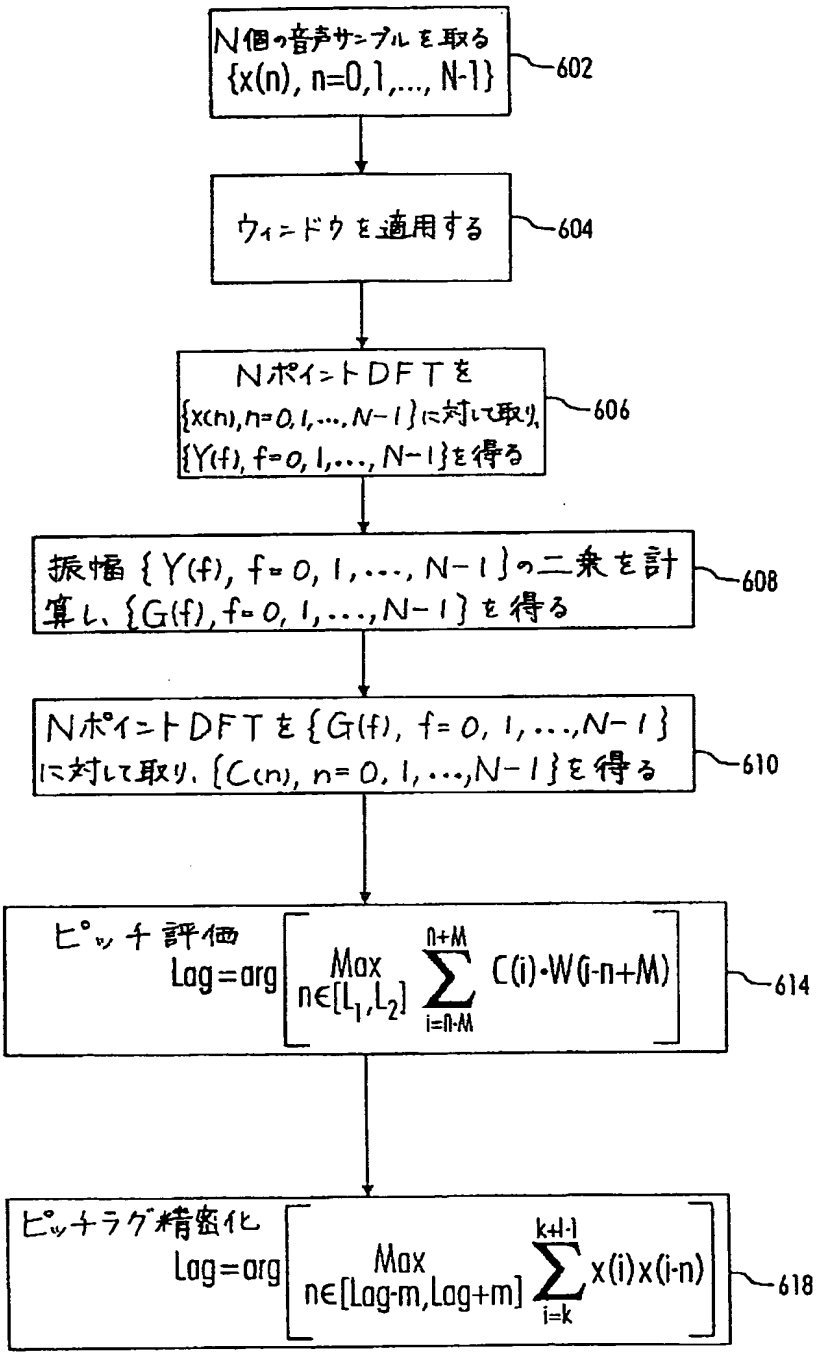
【図８】



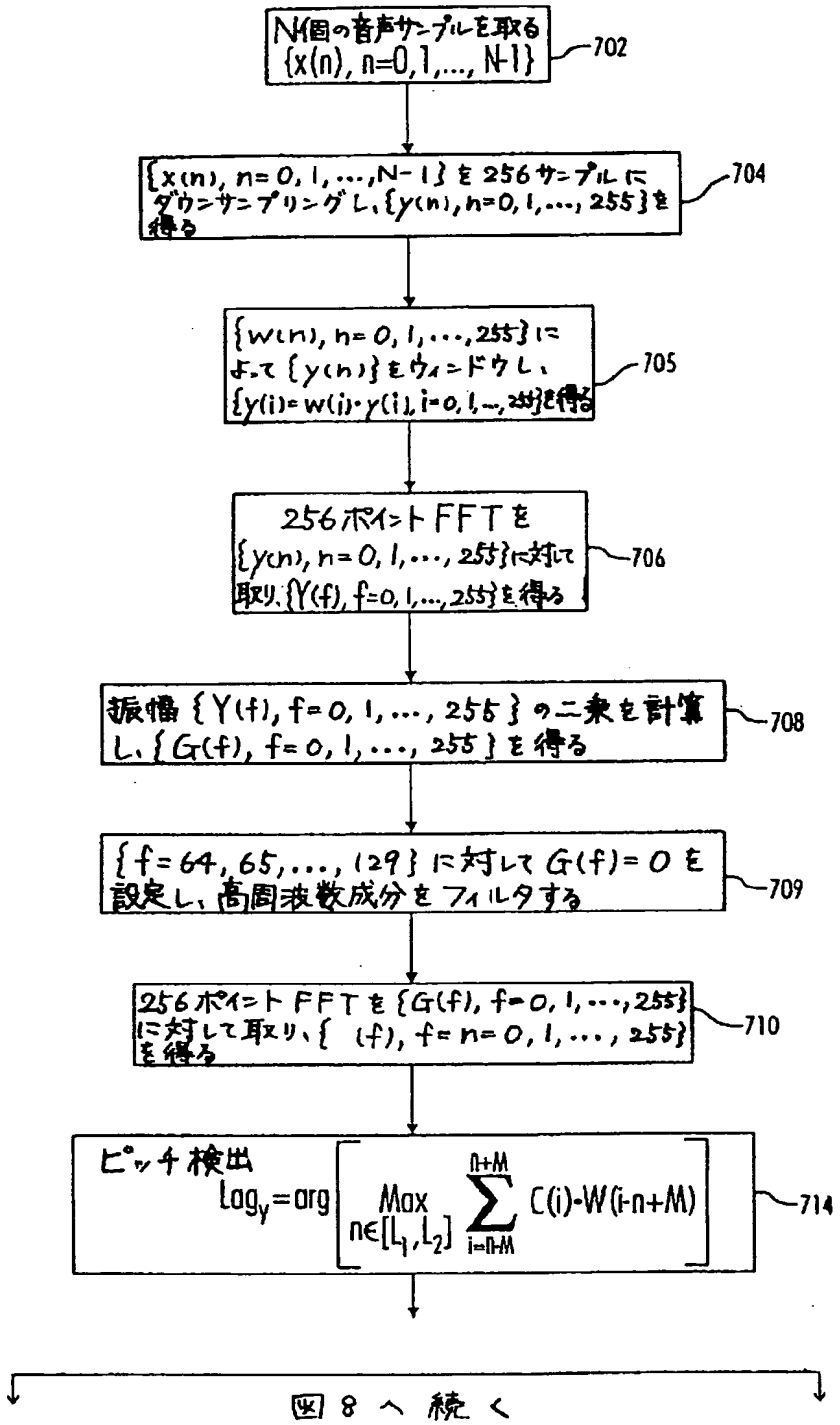
【図９】



【図６】


N個の音声サンプルを取る
{x(n), n=0,1,..., N-1}
602
ウィンドウを適用する
604
Nポイント DFT を
{x(n), n=0, 1, ..., N-1} に対して取り,
{Y(f), f=0, 1, ..., N-1} を得る
606
振幅 {Y(f), f=0, 1, ..., N-1} の二乗を計算し、{G(f), f=0, 1, ..., N-1} を得る
608
Nポイント DFT を {G(f), f=0, 1, ..., N-1} に対して取り、{C(n), n=0, 1, ..., N-1} を得る
610
ピッチ評価
Lag = arg Max n∈[L1,L2] Σ i=n-M to n+M C(i)·W(i-n+M)
614
ピッチラグ精密化
Lag = arg Max n∈[Lag-m, Lag+m] Σ i=k to k+l-1 x(i)x(i-n)
618

【図７】


N個の音声サンプルを取る
{x(n), n=0,1,..., N-1}
702
{x(n), n=0, 1, ..., N-1}を256サンプルにダウンサンプリングし、{y(n), n=0, 1, ..., 255}を得る
704
{w(n), n=0, 1, ..., 255}によって{y(n)}をウィンドウし、{y(i)=w(i)·y(i), i=0, 1, ..., 255}を得る
705
256ポイントFFTを{y(n), n=0, 1, ..., 255}に対して取り、{Y(f), f=0, 1, ..., 255}を得る
706
振幅{Y(f), f=0, 1, ..., 255}の二乗を計算し、{G(f), f=0, 1, ..., 255}を得る
708
{f=64, 65, ..., 129}に対してG(f)=0を設定し、高周波数成分をフィルタする
709
256ポイントFFTを{G(f), f=0, 1, ..., 255}に対して取り、{ (f), f=n=0, 1, ..., 255}を得る
710
ピッチ検出
$\log_y = \arg \left[\max_{n \in [L_1, L_2]} \sum_{i=n-M}^{n+M} C(i) \cdot W(i-n+M) \right]$
714
図８へ続く

フロントページの続き

(72)発明者　トム・ホン・リー
アメリカ合衆国、07748　ニュージャージー州、ミドルタウン、ノウルウッド・ドライブ、501